문서 분류에 이용 가능한 벡터 공간의 확장 방법

이 상 곤, 유 경 석
전주대학교 컴퓨터공학과 언어과학실
e-mail : samuel@jj.ac.kr, rudtjr0094@naver.com

An Expansion of Vector Space for Document Classifications

Samuel Sangkon Lee and Kyungseok Yoo
Dept. of Computer Engineering and Engineering, Jeonju University

요 약

본 논문에서는 한국어 문서의 분류 정밀도 향상을 위해 애매어와 해소어 정보를 이용한 확장된 벡터 공간 모델을 제안하였다. 벡터 공간 모델에 사용된 벡터는 같은 정도의 가중치를 갖는 축이 하나 더 존재하지만, 기존의 방법은 그 축에 아무런 처리가 이루어지지 않았기 때문에 벡터끼리의 비교를 할 때 문제가 발생한다. 같은 가중치를 갖는 축이 되는 단어를 애매어라 정의하고, 단어와 분야 사이의 상호정보량을 계산하여 애매어를 결정하였다. 애매어에 의해 애매성을 해소하는 단어를 해소어라 정의하고, 애매어와 동일한 문서에서 출현하는 단어 중에서 상호정보량을 계산하여 해소어의 세기를 결정하였다. 본 논문에서는 애매어와 해소어를 이용하여 벡터의 차원을 확장하여 문서 분류의 정밀도를 향상시키는 방법을 제안하였다.

1. 서론

분류 기술의 일반적인 방법은 각 분야에 분류되어 있는 문서의 정보를 벡터로 표현한 벡터 공간 모델이 사용된다. 작성된 벡터의 각 축은 단어, 각 축의 가중치는 단어의 출현 빈도값이 사용된다. 실제 문서를 분류할 때에는 새로 입력된 문서에 대해 벡터를 작성하여, 각 분야의 벡터와의 내적에 의해 그 유사도를 계산하고 가장 유사한 분야로 분류된다. 작성된 벡터는 복수 분야의 벡터와 같은 가중치를 갖는 축이 존재한다.

본 논문에서는 여러 분야의 벡터와 같은 가중치를 갖는 축이 되는 단어를 '애매어'라 정의하고, 이를 이용하여 별도의 정보와 함께 벡터를 확장하고, 원래 단어가 가진 애매성을 해소하여 분류의 정확성을 보장하고자 한다. 애매어 모두를 확장하면 그 계산량이 많아지므로 여러 분야를 지칭하는 애매어에 대해서만 벡터를 확장하여 계산량을 감소시킨다. 확장의 대상이 되는 애매어의 추출은 상호정보량[11-15]을 측정하여 그 결과값을 이용하고, 애매성 해소를 위한 정보는 애매어와 함께 동일 문서 내에 출현하는 공기어를 적극 이용한다. 공기어 중에는 한 분야에서만 출현하는 공기어를 '해소어가 될 수 있는 후보어'라 정의하고, 이 해소 후보어를 이용하여 벡터 확장을 한다. 확장 이후의 축은 애매어의 빈도와 공기어와 함께 출현한 애매어의 빈도가 된다. 본 논문에서 제안한 방법의 유용성을 평가하기 위해 언론사에서 제공하는 언론 기사를 이용하여 실험하고 평가한다.

이하 2장에서는 벡터 공간 모델을 이용한 문서 분류 방법, 3장에서는 애매어의 명세 방법, 4장에서는 해소어를 이용한 벡터 차원의 확장에 대해 서술하고, 5장에서는 결론을 서술한다.

2. 벡터 공간 모델에 의한 분류 방법

본 연구의 구현을 위해 문서 자동 분류 시스템의 개요를 설명하면 다음과 같다. 문서 자동 분류는 크게 나누어 세 가지의 처리를 포함하고 있다. 키워드 추출 모듈, 문서 정보의 작성 모듈, 그리고 문서 분류 모듈 등 세 가지이다.

처리의 흐름은 다음과 같다. 먼저 분류 작업에 이용할 분야 정보를 작성하고, 그 위에 새로 입력된 문서에 대한 분류용 문서 데이터를 작성한다. 그것과 이전에 작성된 분야 정보와의 집합을 비교한다. 그 결과를 참조하여 최종 분류를 행하는 순서로 작성한다. 앞에서 언급한 세 가지 처리 중 키워드 추출 모듈과 문서 정보의 작성 모듈은 분류용 문서 데이터를 작성할 때에, 문서 분류 모듈은 문서 정보를 비교하고 분류할 때에 각각 사용된다. 다음에 세 가지 처리에 대해 서술한다. 키워드 추출에서는 문서에서 그 문서의 특징을 가장 잘 설명하는 중요어를 추출한다. 여기서, 중요어와 키워드는 동일한 것으로 간주한다. 키워드 추출은 문헌 검색, 텍스트 편집 등 폭넓은 분야에서 응용되는 기본 기술이다. 현재의 키워드 추출은 크게 나누면 두 가지 방법이 제안되어 있다. 하나는 통제어 방식이고, 다른 하나는 자유어 방식이다. 통제어 방식은 통제어 사전(시소러스)을 사용하는 방식이다. 키워드의 후보어로 가능한 단어를 미리 시소러스 내에 준비하여 두고, 시소러스에

등록된 키워드가 대상 문서 내에 존재하는가에 의해 그 추출 여부가 결정된다[5]. 다른 것은 자유어 방식인데, 이 방법은 시소러스를 사용하지 않고 대상 문서를 형태소 해석하여 그 해석 방법의 기술 수준에 따라 단어를 분할하고 분할된 형태소열에서 키워드 패턴과의 조합이나 빈도 정보 등의 가중치 계산을 통해 키워드를 추출하는 방식이다. 통제어 방식에서 추출 처리는 단순하지만, 관리에 많은 노력이 필요하여 동시에 대량의 시소러스도 필요하다. 반면에 자유어 방식에서 단어의 추출 처리는 다소 복잡하지만 상대적으로 시소러스의 관리가 필요 없어 통제어 방식에 의한 키워드 추출 방법보다 자유어에 의한 방식이 현재에 많이 사용된다.

문서 정보 작성부는 키워드 추출을 이용하여 추출된 문서의 특징을 나타내는 정보를 서로 비교하기 쉽도록 요약하고 각각의 문서 정보를 작성한다. 문서 정보는 문헌 검색, 문서 분류, 요약 문장 생성 등의 과정에서 이용되는 기본적인 데이터이다. 문헌 검색은 대량의 데이터에서 자신이 지정한 정보를 갖는 데이터를 검색하고 선택하는 방법이다. 문서 분류에서 작성된 문서 정보와 미리 작성되어 있는 분류 체계 정보를 비교하여 그 결과가 가장 좋은 쪽으로 분류해 가는 작업이다. 문서 정보를 저장하는 데이터 방식은 전치 인덱스 방법과 벡터 형식의 표현 방법 등이 있다.

문서 분류 모듈은 문서 정보 작성 모듈에서 작성된 분류를 원하는 정보를 특정한 기준에 의해 선별하여 나누는 기술이다. 실제로 분류할 때는 인간이 미리 작성한 분류 정보를 적극적으로 이용한다. 문서 분류 기술은 전자화 된 데이터의 증가에 따라 그 중요성 또한 증가하고 있어 보다 효율적인 기술이 개발되어야 한다.

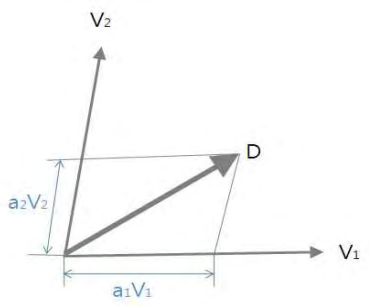
3. 분류 모델에 적용

문서 분류 모듈은 문서 정보 작성 단계에서 작성한 분류하고자 하는 문서 정보를 특정한 기준에 의해 선별하는 기술이다. 분류할 때는 미리 작성된 분류 정보를 이용한다. 인터넷의 많은 정보가 미리 특정한 분류 기준에 따라 정리되어 있으며 개인이 필요한 정보를 즉시 검색하는데 비교적 용이하다. 이와 같이 정보를 분류하는 것은 정보 접근을 지원하는 또 다른 방법이라 생각된다. 이와 같이 문서 분류에 대한 연구는 그 중요성이 증대되고 있다[8].

본 논문에서는 벡터 형식의 문서 정보를 분류 결과에 적극 반영한다. 다음에 그 개념을 설명한다. 전절에서 서술한 바와 같이 어느 문서 데이터의 벡터 정보는 (식 1)로 표시할 수 있다. 분류된 문서 정보에 대해서는 (식 2)와 동일한 벡터 정보를 이용할 수 있다. 문서의 벡터 정보는 아래의 식으로 표시할 수 있다.

$$D = \sum_{i=1}^{t} a_i V_i \cdots\cdots\cdots\cdots\cdots\cdots \text{(식 1)}$$

$$Q = \sum_{i=1}^{t} q_i V_i \cdots\cdots\cdots\cdots\cdots\cdots \text{(식 2)}$$



(그림 1) 문서의 벡터 표현

위의 식에서 계수 q_i 는 앞의 식 a 와 같은 것이다. 여기서 가장 단순한 경우에 검색 질문에 대해 색인어 T_i 가 존재하는 경우는 1, 존재하지 않는 경우는 0 이 된다. 보다 복잡한 경우는 분류된 문서 Q 에 대해 색인어 T_i 의 중요도는 표시하는 값이 입력된다. 예를 들면, 출현 빈도나 그 것을 전체 빈도로 나눈 정규화 된 값이다. (그림 1)에 2차원 벡터 공간에서의 문서 표현의 예를 나타낸다.

이와 같이 본 연구에서는 문서를 벡터의 선형 결합을 이용하여 그 의미를 표현하였으며, 이에 의해 문서 사이의 비교나 분류 작업이 벡터 연산으로 가능함을 입증한다. 벡터 공간에서 두 가지 벡터의 유사도는 여러 형태로 정의되지만 본 연구는 두 가지 벡터가 이루는 각의 코사인 값을 이용한다.

$$x \cdot y = |x|\,|y| \cos\alpha \cdots\cdots\cdots\cdots\cdots \text{(식 3)}$$

여기서 $|x|$는 벡터의 길이를 나타내고, α는 벡터가 이루는 각을 나타낸다. 이러한 유사도를 이용하는 경우 문서 정보 D 와 질의어 Q의 유사도는 이하의 식과 같다.

$$sim(D, Q) = D \cdot Q = \sum_{i,j=1}^{t} a_i Q_j V_i \cdot V_j \quad \text{(식 4)}$$

여기서 간단히 하기 위해 t 개의 키워드에 대응한 벡터 V 는 각각 직교한다고 가정한다. 이 때,

$$V_i \cdot V_j = \begin{cases} 0, & i \neq j \text{일 때} \\ 1, & i = j \text{일 때} \end{cases} \quad \text{이다.}$$

따라서 $sim(D, Q)$는 다음 식에 의해 간략화 된다.

$$sim(D, Q) = \sum_{i=1}^{t} a_i q_i \cdots\cdots\cdots\cdots \text{(식 5)}$$

그러나 내적은 벡터의 크기에 의해 영향을 받으므로 두 가지 벡터가 이루는 각의 코사인 값을 이용하여 이 내적에 기초한 유사도를 정규화 한다. 즉, 내적이 아니고 두 개의 벡터가 이루는 각의 코사인 값을 취한다. 정규화 한 유사도를 $sim'(D, Q)$라 하면 이하의 식과 같이 표시할 수 있다.

$$sim'(D, Q) = \frac{sim(D, Q)}{|D|\,|Q|} \cdots\cdots\cdots \text{(식 6)}$$

이와 같이 문서 분류 연구에서 벡터를 사용하면 분류 정보와 문서 정보를 비교하는 유사도를 정의할 수 있다. 또한 코사인 유사도는 문서 데이터의 벡터 Q와 분류 정보의 벡터 D 가 이루는 각이 적을수록 두 가지 정보가 유사하게 된다.

앞에서 서술한 두 종류의 문서 분류 방법에서는 벡터 정보를 어떻게 활용하여 문서를 분류할 것인가에 대해 서술하였다. 분류 기술의 개념도를 아래의 (그림 2)에 제시

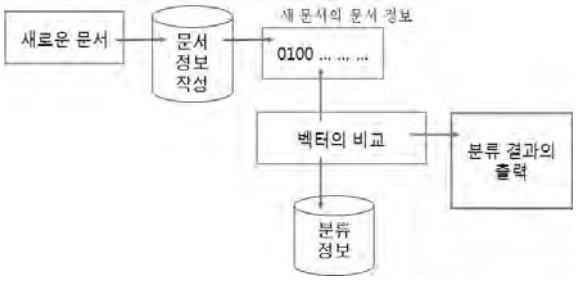
<표 3> 동일한 가중치를 갖는 벡터

	단어1	단어2
문서1	*10*	*10*
[분야1]	*10*	*5*
[분야2]	*10*	*15*

<표 4> 상호정보량의 계산

	단어 1	단어 2
분야 1	10	5
분야 2	10	15

하였다.



(그림 2) 문서 분류기의 개념도

문서 분류에서 사용되는 방법은 인간이 미리 분류하여 작성한 데이터를 기본으로 분류 데이터를 작성하고, 새로 입력된 문서는 그 문서 정보를 작성하여 분류 데이터와의 유사도를 계산한다. 그 값이 높을수록 해당 분야에 가깝다고 할 수 있다.

4. 구현

본 연구의 구현을 위해 먼저 사용한 개발 모형은 에자일(Agile) 방법론[4, 5]이다. 이 방법론은 프로세스, 도구보다 의사소통을 강조함으로써 요구 사항 변경에 적극적으로 대처할 수 있도록 하는 방법론이다. 문서화 된 계획보다 단위 모듈 개발 결과와 고객의 신속한 요구사항 반영, 그리고 테스트에 중점을 두고 있다. 유스케이스 식별은 세 가지 기능을 구현하였으며, 유스케이스 다이어그램을 제시하였다. 식별 작업은 요구 분석 -> 설계 -> 개발 -> 테스트의 순으로 작성한다. 이 세 가지 기능의 클래스 다이어그램을, 시퀀스 다이어그램을, 그리고 유스케이스 명세서를 제공하여야 한다.

5. 결론

본 논문에서는 애매어와 해소어의 상호정보량을 벡터 공간 모델에 적용하여 문서 분류의 정밀도 향상에 관해 연구하였다. 벡터 공간 모델에 사용된 벡터는 같은 정도의 가중치를 갖는 축이 하나 더 존재하지만, 기존의 방법은 그 축에 아무런 처리가 이루어지지 않았기 때문에 벡터끼리의 비교를 할 때 문제가 발생한다.

같은 가중치를 갖는 축이 되는 단어를 애매어라 정의하고 단어와 분야 사이의 상호정보량을 계산하여 애매어를 결정하였다. 애매어를 갖는 애매성을 해소하는 단어를 해소어라 정의하고 애매어와 동일한 문서에서 출현하는 단어 중에서 상호정보량을 계산하여 결정하였다. 해소어에 의해 해소되는 애매어를 갖는 정보를 벡터 비교에 반영시키기 위해 애매어의 축을 “애매어”와 “해소어와 동시에 출현하는 애매어”라 하여 벡터의 차원을 확장하여 분류 정밀도를 향상시키는 방법을 제안하였다. 인터넷 포털사이트 네이버의 뉴스 사이트에서 언론사별 뉴스를 수집하여 실험을 하고 그 유효성을 확인하였다.

향후의 연구로서는 애매성을 갖고 있는 분야의 정보 뿐 아니라 모든 분야의 정보를 고려한 해소어를 결정하는 방법에 대해 연구하고자 한다. 본 논문의 방법은 계산 속도를 고려하지 않았으므로 미래에는 처리의 고속화에 대해 연구하고자 한다.

참고문헌

[1] 정경희, “의학 분야 웹 자료의 분류에 대한 개선 방안 연구”, 정보관리학회지, 제21권, 제2호, pp. 089-106, 2004.

[2] 윤성희, 백선욱, “단어 의미 정보를 활용하는 이용자 자연어 질의 유형의 효율적 분류”, 정보관리학회지, 제21권, 제4호, pp. 251-263, 2004.

[3] 이원휘, “K-Means 알고리즘을 이용한 대용량 문서 클러스트링에서 개선된 초기 중심 선정 방법의 제안”, 전북대학교 대학원 컴퓨터공학과 박사학위 논문, pp. 1-101, 2010.

[4] 안동언 외, 최신 정보검색론, 교보문고, pp. 1-514, 2010.

[5] 이상곤 외, "개념 기반 복합 키워드 추출 방법", 한국컴퓨터교육학회 논문지, 제6권, 제2호, pp. 23-31, 2003.

[6] 이상곤, "한글 문서 분류용으로 이용할 복합어로 구성된 분야 연상어의 추출법," 정보과학회 논문지: 소프트웨어 및 응용, 제32권, 제7호, pp. 636-649, 2005.

[7] 노대욱 외, "정보 검색 기술을 이용한 비지도 학습 기반 문서 분류 시스템 개발," 정보과학회논문지 : 소프트웨어 및 응용, 제34권, 제2호, pp. 123-130, 2007.

[8] 양재군, 배재학, 이종혁, "온톨러지 재사용을 위한 범주 재분류", 정보처리학회논문지(B), 제12권, 제1호, pp. 69-80, 2005.